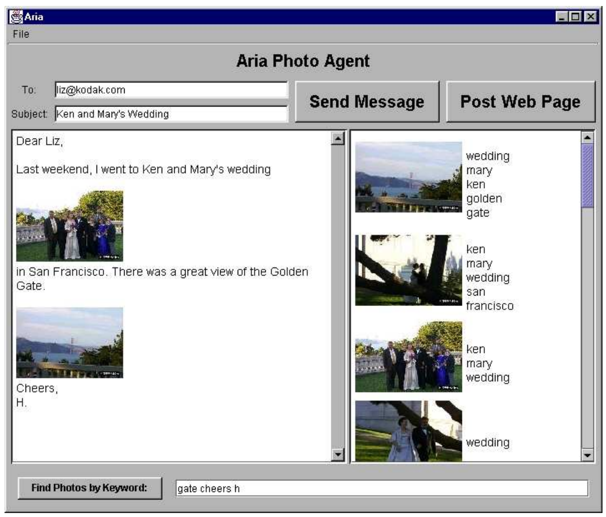
An Agent for Integrated Annotation and Retrieval of Images

Henry Lieberman
Media Laboratory
Massachusetts Institute of Technology
Cambridge, Mass. 02139 USA
(+1-617) 253-0315
lieber@media.mit.edu

Elizabeth Rosenzweig
Eastman Kodak Company
Research Labs
Lowell, Mass. 01851 USA
(+1-978) 323-7513
erosenz@kodak.com

Push Singh
Media Laboratory
Massachusetts Institute of Technology
Cambridge, Mass. 02139 USA
(+1-617) 253-1750
push@media.mit.edu



ABSTRACT

With the advent of digital photography, consumers now accumulate a large number of images over their lifetime. These images are often stored in "shoeboxes", rarely looked at, occasionally put into albums, but usually laying around, unused and unlooked at for years. Until computer vision reaches the point where images can be automatically analyzed, most image retrieval will depend on textual keywords attached to specific images. But annotating images with keywords is a tedious task, and, with current interfaces, ordinary people cannot reasonably be expected to put in the upfront effort to annotate all their images in the hopes of facilitating future retrieval. Furthermore, retrieval applications themselves are awkward enough that they often go unused in cases where the user might indeed find images from the library useful.

Part of the problem comes from the conventional view that annotation and retrieval are two completely separate operations, to be addressed by applications operating independently from each other, and from any application in which the images might be used.

We present a user interface agent, ARIA [Annotation and Retrieval Integration Agent], that assists users by proactively looking for opportunities for image annotation and image retrieval in the context of the user's everyday work. The initial version of ARIA sits in the user's e-mail editor, and continuously monitors typing. Continuous, ranked searches are automatically performed from an image library, and images relevant to the current text can be inserted in a single click. Descriptions of images entered for the purposes of storytelling in e-mail can be seamlessly employed as raw material for image annotation. While not completely automating the image annotation and retrieval process, ARIA dramatically reduces user interface overhead, leading to better annotated image libraries and fewer missed opportunities for image use.

Keywords

Agents, image databases, image retrieval, image annotation, image management, metadata

1. You push the button, we do the rest

George Eastman s original advertising slogan for Eastman Kodak Company was You push the button, we do the rest . What he sought was to convince the consumer that the technology of photography, including Kodak s products and services, would act *as an agent* for the consumer in recording the memories of their lives. Ironically, in those days, that meant shooting film in a sealed camera, and mailing the camera back to Kodak for processing.

Modern photography, especially digital photography, has come a long way, but the process of making and effectively using photographs still is much more tedious than it should be. We aim to reduce this tedium by using software agents rather than human labor, as much as possible, to implement the we do the rest part. Furthermore, the rest doesn t end when a finished hardcopy picture reaches the hands of the consumer, but also includes any future use that the user might make of the picture.

We are particularly interested in the shoebox problem , because it is an untapped source of communicating shared memories that is currently lost. After initially viewing pictures (after they are returned from film developing or downloaded to a computer), many people accumulate their images in large archival collections. In the case of hardcopy photos or printouts, often in conveniently-sized shoeboxes or albums. Images in shoeboxes, or their electronic equivalent in folders or removable media, are often never (or very rarely seen again), because of the difficulty of retrieving specific images, browsing unmanageably large collections and organizing them. Typically, any organizing apart from rough reverse-chronological order involves so much effort on the part of the user that it is usually never performed.

Potentially, the images could be annotated with text labels and stored in a relational database and retrieved by keyword. However, consumers cannot ordinarily be expected to put large amounts of upfront effort into classifying and categorizing images by keyword in the hopes of facilitating future retrieval. User testing shows that they won t do it in practice. Furthermore, the retrieval itself involves dealing with a search engine or other application that itself imposes overhead on the process, even if only the overhead of starting and exiting the application and entering keywords. Because of this overhead, opportunities to use images are often overlooked or ignored.

Further down the road, one could imagine that automated image analysis could identify people, places, and things in a photograph and do image annotation automatically. Though considerable progress has been made in this area [2, 4, 10] we are still far short of being able to rely on this approach. In addition, even if the image can be interpreted, many salient features of images exist only in the user s mind and need to be communicated somehow to the machine in order to index the image. Therefore, retrieval, based on textual annotation of images, will remain important for the foreseeable future.

We see a role for a user interface agent in facilitating, rather than fully automating, the textual annotation and retrieval process. The role of the agent lies not so much in automatically performing the annotation and retrieval but in *detecting opportunities* for annotation and retrieval and alerting the user to those opportunities. The agent should also make it as easy as possible for the user to complete the operations when appropriate.

2. A scenario: Telling the story of a wedding

A primary use of consumer picture-taking is connecting people through pictures and stories they tell about events. Pictures convey emotions in a way that words cannot. Imagine you have recently attended a wedding, and are sending an electronic mail message to a friend describing the event. The mail would be greatly enhanced if you could illustrate the story by including pictures of the event, and perhaps also of pictures of related people, places, and events in the past. What do you need to do to accomplish this?

- Take pictures at significant events in the wedding: exchanging vows, cutting the cake, the couple kissing, etc. Take pictures at each dinner table, people dancing, conversing, etc. For the moment, we won t go into details about the complications of image capture: e.g., setting exposure modes, running out of battery power, filling up the storage medium, etc.
- Get the pictures into the computer. This might involve: Removing the storage medium [memory card, floppy disk] from the camera and inserting it into a reader. Possibly connecting the reader device or the camera with a cable to the computer. Launching the communications

software or setting a mode to perform the transfer. Selecting a place on the computer for the pictures to go. Selecting a name for the set of pictures so you don t forget what they are.

- Launching an image viewing/manipulation/cataloging program [e.g., Adobe Photoshop, PicturePage]. Initially scanning the pictures and removing the duds , exposures that were not successful. Possibly changing the file name of an individual picture to describe its contents. If you do have an image database, you may attach keywords to individual images or sets. Possibly performing image manipulation on the picture [cropping, adjusting brightness, etc.] using the same or separate application. Possibly printing hardcopy of images for storage or sending to others. Possibly e-mailing pictures to others or posting on Web pages.
- Perhaps weeks or months later, you would like to use the images when composing an e-mail message to a friend or family member about the wedding. In addition to launching and operating the e-mail application itself, you must launch another application, an image viewer/catalog/search application. Perhaps you may search around in the file system to find a folder containing relevant images, either by browsing or retrieval by file name. Perhaps relevant images are stored on your own or acquaintances Web pages, necessitating launching the Web browser and typing URLs or using search engines. Perhaps you may search the image database via keywords. You switch back and forth between applications as the need arises. If you succeed in finding a picture you cut the picture from the image application and paste it into the e-mail editor.
- Nothing about this interaction would make this task easier to do the next time, for example, if you wanted to tell a related story to a different person in the future.

You push the button, you do the rest.

3. Integration of annotation, retrieval, and use: The importance of context

Part of the problem comes from the mistake of thinking about annotation, retrieval, and use of images as each being the province of a separate application . This leaves the burden on the user to enter and leave applications when appropriate, and explicitly transfer data from one application to another, usually via cut and paste. Users are inclined to think about their own tasks, as opposed to applications and data transfer. Each user s task, such as sending an e-mail message, carries with it a context, including data being worked with, tools available, goals, etc.

Our approach is to try to integrate image annotation, retrieval, and use into a single application . Following a user-centered design process, it s the use that is most important, and we picked text editing, in particular editing e-mail messages, as an application for which picture retrieval from an image library might naturally find use. We integrated a retrieval and annotation application to eliminate the confusing context switch of having separate applications.

Much of what we call intelligence in problem solving is really the ability to identify what is relevant and important in the current problem solving context and make it available for use in a just-in-time manner [8]. The job of the agent in the integrated application is to make available and conveniently accessible the appropriate context for relating text and images.

For the task of editing e-mail messages, typing text descriptions often sets up a semantic context in which retrieval of pictures relevant to that context would be appropriate. Similarly, seeing the pictures sets up a context for which some textual descriptions present in the current context may apply, setting up an opportunity for annotation. We have designed the interface to make taking advantage of these opportunities, in both directions, only a single click away.

4. ARIA: Annotation and Retrieval Integration Agent

ARIA [Annotation and Retrieval Integration Agent] is the prototype integrated annotation and retrieval application we have implemented to test some of these ideas. The initial implementation consists of the standard Java Swing text editor, coupled to a pane containing a custom-built image retrieval and annotation application. At the bottom of the text editor, a retrieval term is displayed, taken from the text surrounding the editor s cursor. To the right of the text editor, a column of retrieved images is displayed. This column is dynamically updated. To the right of each image, a list of annotation keywords pertaining to that image appears. The screen configuration appears in Figure 1.

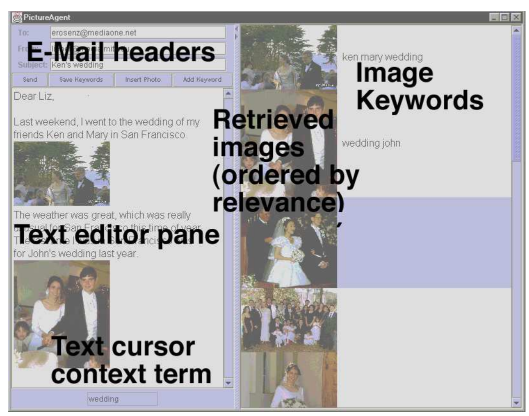


Figure 1. ARIA screen layout

The user can freely type in the text editor, and the text editing actions are recorded by the ARIA agent. The agent is continuously running, and observes the user s typing actions in the text editor, as in the Remembrance Agent [11], which also worked in the domain of text editing, and Letizia [6] which observed URLs chosen by the user in the course of Web browsing.

Input to the agent is analyzed to extract keywords in the context surrounding the text cursor. We are experimenting with different methods of performing the text analysis, but a straightforward approach is to use common information extraction heuristics such as TFIDF 12, common word elimination, etc. Keywords in a neighborhood of the cursor are continuously displayed in the box below the text editor pane. The keywords are used to query the image database, and a ranked list of pictures is graphically displayed in the column to the right, in order of relevance. This list is recomputed at every keystroke.

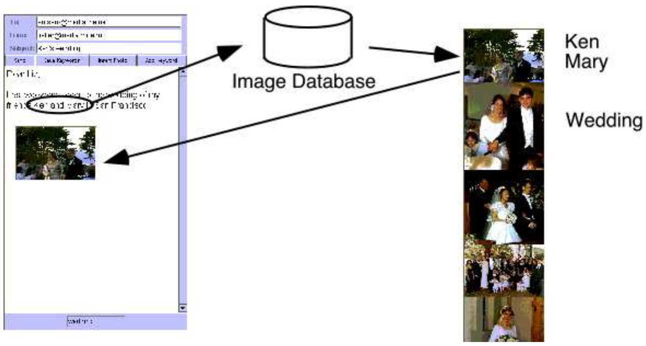


Figure 2. Image retrieval in ARIA

For example, we start by typing into the text editor,

> *Dear Liz,*
>
> *Last weekend, I went to the wedding of my friends Ken and Mary in San Francisco .*

and the agent extracts the keywords Ken , Mary , wedding , friends , and San Francisco . Assuming, for the moment, that the images happened to have been previously annotated with these terms, this would be sufficient to retrieve pictures of the event, which would appear in the column to the right of the text editor. We are likely to see a picture that would make a nice illustration for the e-mail message at this point, and a single drag inserts the picture into the editor.

Notice how different this interaction is from conventional image retrieval.

- We didn t switch to an image retrieval application.
- No keyword queries were typed.
- No file dialog boxes.
- No cut and paste was needed to use the image.

Total user interaction: one drag.

Of course, the desired image might not appear immediately, in which case we could scroll through the list until we found a suitable image, call up a dialog box to load other image sets, etc. Even in that case, we d still be saving some interaction compared to the conventional approach, but what s important is that the most likely case is put right at our fingertips, so the average interaction would be much reduced.

We are also experimenting with using other kinds of information that appear in the typed text to aid retrieval. One is temporal references. Typing "Last weekend " leads the system to look up the date, compare it to the dates timestamped on every picture, and retrieve the pictures that have these dates. We will include a large vocabulary of time references, including specific and relative dates, intervals, "about", etc.

But suppose the desired image had not yet been annotated, as would be the case if we were loading these images for the first time, for example, by inserting the image storage media [in the case of our camera, a Kodak DC265 zoom digital camera, a compact flash card].

As soon as you insert the flash card media into the computer, ARIA *immediately* loads the images on the flash cards, without any further user action.. ARIA is set up to automatically poll for incoming input every few seconds. This may seem like a small thing, but it removes the need for

- Invoking a "Load Pictures" operation
- Waiting for a file dialog box to pop up
- Figuring out where the pictures should go in the file system, and what they should be called [our camera simply names them with a meaningless name, e.g. P000007.jpg]
- Remembering the names so that you can retrieve them next time.

The pictures just inserted are immediately brought to the top of the retrieval window, since it is likely the user will want to see them immediately. Scrolling through and selecting an image to illustrate the opening sentence of the letter typed above, we note that the image now appears sitting in the editor next to text containing many keywords appropriate for describing it.

This represents an opportunity to use the existing text to annotate the images so that retrieval will be easier next time. Aria incorporates an automatic annotation feature that uses keywords extracted from the text surrounding the image [sometimes descriptions appear before, sometimes after the image, but usually close] to annotate that image.

For example, if we type

> *Here's a great view of the Golden Gate*

Aria places the annotations "Golden Gate" on the picture inserted. The user types the description of the picture in order to communicate to another person, not specifically to annotate the image, but once the computer has the input, why not take advantage of it? This *repurposing* of user input is an important aspect of agents in general, and a key to reducing the burden on the user. ARIA's guesses are, of course, not perfect, and we also leave to the user the option of editing out incorrect guesses. If you click on an image, ARIA displays a keyword editor, so that the user can select a set of appropriate keywords at once, and avoid the irrelevant ones

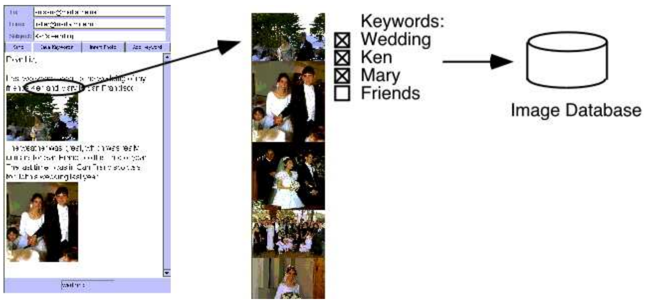


Figure 3. Image annotation in ARIA

We can also manually drag words from the text editor to an image to place an annotation on that image. Sometimes an appropriate annotation is missed by ARIA, or appears too far away from the image to be considered for annotation. Again, the interaction is streamlined relative to what it would be in a conventional image annotation application, which would require typing, cut and paste, or selection from an external list of keywords, as well as a context-switch in and out of the annotation application.

When the user is finished telling the story, they can either "Send Message", or "Post Web Page", the two most common scenarios for using images. Each of these actions is accomplished by pressing a single button.

Figure 4. Web page produced by ARIA

That's it!

5. User Testing

User studies of Aria are underway at Kodak s Boston Software Development Center in Lowell, Mass. At submission time (August 2000) we have only run a small number of participants (4), but the studies included in-depth interviews and observations with each participant (over three hours for each participant, during two sessions). We feel that our preliminary results are indicative of what would be found in a larger study.

Each participant came with a memory card full of photographs or was given a card by the experimenters. They were asked to load the pictures and compose an e-mail message to a friend using at least three of the pictures. They were asked to do the task, first with Aria, then with Eastman Kodak Company s Picture Easy software, or vice versa. Picture Easy is a conventional image editing and cataloging application that ships with Kodak cameras. Participants were given a brief demo of each program. All participants had used digital cameras and e-mail, but were not computer professionals or programmers.

During the first session, in addition to the e-mailing task, participants were told they could do whatever organizing activities they thought might help them find pictures in the future (annotations, naming, folders, albums, etc), though they were not required to do anything but send the e-mail message. Two weeks later, the participants were brought back, and told to write a letter to a different person about the same event. We wanted to see what they would choose to do, and whether Aria s annotations or conventional albums or folders would be helpful in finding photos or remembering story details they may have forgotten after the two week hiatus.

Participants loved Aria s automatic loading of images -- Couldn t be easier , Piece of cake . Participants described the process of selecting pictures and e-mailing messages with Aria as quick, fun and easy. In particular, participants liked incorporating pictures in-place into the text: "It was a big surprise to me that you could click and drag a photo into the message."

Participants particularly liked that they were able to view their pictures while writing their e-mail message, without having to switch applications or modes. One subject did observe, that in the case of just sending pictures without any text at all, which he often did, that it might be faster to use conventional attachments .

Aria s automatic image retrieval surprised and delighted the participants, one exclaiming Oh wow, way cool! when this feature first appeared. In the second test, as we had hoped, Aria automatically brought up appropriate pictures as users typed, making access to their previously annotated images much easier, and serving as useful reminders of the previous story.

Aria served as a useful reminder in both directions -- from the story to the pictures and from the pictures to the story. Storytelling with pictures was an iterative process -- a detail of the story would bring up an appropriate picture, which then triggered more memories of the story in the user's mind, and so on.

Phase I

	Picture Easy	Aria
It was easy to learn how to use the program.	5.5 (1.9)	6.5 (0.6)
I could effectively use the program.	5.8 (1.5)	6.3 (1.0)
I could efficiently complete tasks using the program.	6.0 (1.4)	6.5 (1.0)
Overall, I am satisfied with the program and would use it to email pictures.	5.0 (2.2)	6.5 (1.4)

Phase II

	Picture Easy	Aria
I could effectively use the program.	4.5 (0.9)	6.3 (1.2)
I could efficiently complete tasks using the program.	4.3 (1.5)	6.0 (1.0)
Overall, I am satisfied with the program and would use it to email pictures.	4.0 (2.0)	5.5 (0.5)

Aria s automatic annotation and retrieval was well appreciated by the users. As one user put it, Otherwise, you could go through all the files in your C: drive and still not find the pictures you were looking for .

In contrast, none of the features of Picture Easy actually served to aid retrieval or act as reminders during the second test, other than simply browsing the contact sheet of thumbnail pictures. Though several users initially bemoaned the lack of folders, albums, or any other grouping mechanism in Aria, only one user actually created an album in Picture Easy, which he named, unhelpfully, Story JA [JA were his initials], probably not much use after a long time period had passed. No users created folders in the file system, moved any of the files into existing folders, nor renamed any of the files from their camera-supplied meaningless names [e.g. P000007.jpg] during the test.. Several expressed guilt that they hadn t had the time to organize their home photo collections into folders or properly name files.

Some expressed reservation that Aria might, in some cases, annotate or retrieve the wrong things. Some annotations that Aria proposed weren t correct, but having a few incorrect annotations didn t seem to hurt things, especially when compared to the prospect of having only little or no user-supplied annotation. Users could edit Aria-supplied annotations to remove incorrect guesses, and some did. We didn t observe any egregious cases of mislabeling in the test, but the overall accuracy of annotations would be best assessed only after long-term use. Most negative comments on Aria concerned the lack of features in our prototype, built using Java e-mail and image components, compared to more full-featured applications such as Eudora or Photoshop -- Spell checkers, thesauri, resizable thumbnails and contact sheets, image editing, etc.

The participants also gave us some helpful suggestions, such as the need to maintain consistency between the annotations and text even in the case the text is subsequently edited. They also wanted to be able to go from a picture to a set of past e-mail messages that contained that picture.

Results of a summary questionnaire showed that Aria scored decisively better overall than Picture Easy. (Numbers in the table above are on a scale of 1 (strongly disagree) to 7 (strongly agree)). Aria s scores ranged from 5.5 to 6.5, while Picture Easy scored only 4.0 to 6.0.

We had, however, expected to get an even more dramatic difference between Aria and Picture Easy. One surprising problem we hadn't expected in the testing is that it is actually difficult to get users to express frustration about bad software!

The problem is that people are so acclimated to the shortcomings of conventional software that they cease to question them or complain about them. So when an innovative new piece of software eliminates those frustrations you see the difference between "OK" and "Great", not between "Terrible" and "Great". Nobody thinks to complain about having to fumble a search through a hard disk file system with a standard file dialog box, because we are all forced to do it so often. When Aria eliminated the file dialog box for loading pictures, people complimented it, but failed to criticize Picture Easy for having required it in the first place.

An extreme example occurred when Picture Easy completely lost the text of an e-mail message a user was typing. This happened because Picture Easy requires that one choose pictures before starting to type a message, and if the user returns to the picture selection screen, any previously typed text is lost without warning. We were shocked to watch the user calmly say, "Oh, it lost it. I guess I have to retype it" and fail to strongly criticize Picture Easy for this during the evaluation questionnaire. He obviously expected computer software to be unreliable, so nothing seemed unusual!

6. Future Work

Future work will center on taking advantage of more opportunities to use context to determine appropriate situations for image annotation, image library browsing, and retrieval. Perhaps in the future, GPS systems in cameras could even report the location, at which the picture is taken, and locations correlated with map information and locations mentioned in the text, e.g., San Francisco in our example.

We re often asked how this approach will scale up to large image collections. We have some iniital ideas that need to be worked out, but were are also investigating ways of extending the scope of an annotation to more than one image, and automatically annotating groups of images. For example, if one picture is about a wedding, there s a good chance that subsequent pictures taken within a three-hour span and close to the same location, also describe the same event.

Keywords could be related to ontologies and knowledge bases such as WordNet [1] or CYC to do inheritance on keywords or simple inferences. ARIA s retrieval treats the set of images as an unstructured database, but perhaps a better view is to view sets of pictures as a linked network of relations. ARIA s current retrieval is most analogous to the Remembrance Agent s [11], but perhaps a better model is Letizia s [6] incremental breadth-first traversal of neighborhoods of Web pages. Retrieval could search a neighborhood of who-what-where-when-why links and semantic networks of keyword concepts.

In the long run, we are interested in capturing and using common sense knowledge about typical picture-taking situations such as birthday parties or vacation travel to do intelligent retrieval and automatic annotation of images.

Though full-image understanding remains out of reach, image-based retrieval continues to progress. Image retrieval systems based on computable image properties, such as color histograms or textures appear to be achieving some success [2, 4, 10]. Future work might hook up one of these systems to our agent, and allow automatically propagating user-annotated keyword candidates to similar images.

7. Related Work

There are so many image retrieval and annotation applications that exhaustive comparison here is impossible. We believe ARIA is unique in integrating annotation and retrieval with image use, and repurposing the user s text editor input to provide real-time image recommendations and annotations.

The state-of—the-art in consumer-oriented image annotation and retrieval systems is probably best represented by Kuchinsky, et al. [5]. This system does propose some annotation/retrieval integration, but the annotation/retrieval process is still largely divorced from the contextual use of the images in an application like e-mail. Kuchinsky s system does incorporate some automatic analysis of images to propose annotations. However, the system does not do any observational learning from the user as does ARIA. Budzik and Hammond's Watson [3] also has an agent that does observe user actions and perform image retrieval, but does not consider annotation.

ARIA s approach of using a proactive agent for providing real-time suggestions was inspired largely from the use of this approach by the Remembrance Agent [11] in the text editor domain, and Letizia [6] in the Web browser domain.

8. Acknowledgments

The authors would like to acknowledge Charles Judice and Mark Wood at Kodak for their support of this work, Mona Patel and Kristen Joffre for the user evaluation, and Brad Rhodes for his contributions to the code. We would also like to acknowledge the support of the News in the Future and Digital Life Consortiums at the MIT Media Lab.

9. REFERENCES

1. Y. Alp Aslandogan, Chuck Thier, Clement T.Yu, Jon Zou, and Naphtali Rishe, Using semantic contents and WordNet in image retrieval Proceedings of the 20th annual international ACM SIGIR conference on Research and Development in information retrieval, 1997, Pages 286 — 295.
2. Jonathan Ashley, Myron Flickner, James Hafner, Denis Lee, Wayne Niblack, and Dragutin Petkovic, The query by image content (QBIC) system, Proceedings of the 1995 ACM SIGMOD International conference on Management of data, 1995, Page 475.
3. Budzik, Jay, and Kristian J. Hammond, User Interactions with Everyday Applications as Context for Just-in-time Information Access, ACM Conference on Intelligent User Interfaces (IUI-2000), New Orleans, January 2000.
4. S-F. Chang, "Content-Based Indexing and Retrieval of Visual Information," IEEE Signal Processing Magazine, July 1997.
5. Allan Kuchinsky, Celine Pering, Michael L. Creech, Dennis Freeze, Bill Serra, and Jacek Gwizdka, FotoFile: a consumer multimedia organization and retrieval system, ACM Conference on Human-Computer Interface, [CHI-99] Pages 496 — 503, Pittsburgh, May 1999.
6. Henry Lieberman, Autonomous Interface Agents, ACM Conference on Human-Computer Interface, [CHI-97], Atlanta, March 1997.
7. Henry Lieberman, Integrating User Interface Agents with Conventional Applications, Knowledge-Based Systems Journal, Elsevier, Volume 11, Number 1, September 1998, pp. 15-24. ACM Conference on Intelligent User Interfaces, San Francisco, January 1998.
8. Henry Lieberman and Ted Selker, Out of Context: Computer Systems that Adapt to, and Learn from Context, IBM Systems Journal, Vol 39, Nos 3&4, pp.617-631, 2000.
9. Henry Lieberman and Cindy Mason, Intelligent Agents for Medicine, in Future of Health Technology, Renata Bushko, ed. 2001. .
10. A. Pentland, R. Picard, and S. Sclaroff, Photobook: Tools for Content-Based Manipulation of Image Databases, SPIE Storage and Retrieval of Image & Video Databases II, Feb 1994.
11. Bradley Rhodes and Thad Starner, The Remembrance Agent: A continuously running automated information retrieval system, The Proceedings of The First International Conference on The Practical Application of Intelligent Agents and Multi Agent Technology (PAAM '96), London, UK, April 1996, pp. 487-495.
12. G. Salton, Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison Wesley, 1989.
13. Sam Pack, The American Family: Their photographs, and the Stories they tell. An Ethnographic view,

Doctoral Thesis, Temple University, Department of Anthropology, July 11, 1999.

____